JOSEPH: PHONETIC-AWARE SPEAKER EMBEDDING FOR FAR-FIELD SPEAKER VERIFICATION

Zezhong Jin*, Youzhi Tu*, and Man-Wai Mak†
* † The Hong Kong Polytechnic University
Department of Electrical and Electronic Engineering
Hong Kong SAR

***Abstract*—Performing speaker verification (SV) at a distance from the sound source is challenging because of the interference of noise and reverberation. In such a situation, incorporating phonetic information into speaker embeddings can help reduce the adverse effects of noise and reverberation. Inspired by this observation, we propose a *J*ointly *o*ptimized *s*peaker-*e*mbedding and *ph*onetic-matching (Joseph) framework to exploit phonetic content for far-field SV. The framework encourages the speaker embeddings to preserve phonetic information by matching the frame-based feature maps of a speaker embedding network with wav2vec's vectors. The intuition is that phonetic information can preserve low-level acoustic dynamics with speaker information and thus partly compensate for the degradation due to noise and reverberation. Results show that the proposed framework outperforms the standard speaker embedding on the VOiCES Challenge 2019 evaluation set and the VoxCeleb1 test set. This indicates that leveraging phonetic information under far-field conditions is effective for learning robust speaker representations.**

***Index Terms*—Far-field speaker verification, multi-task learning, phonetic content, wav2vec**

I. Introduction

Speaker verification (SV) plays an important role in various fields, such as biometric authentication, e-banking, and access control. Traditional SV models rely on statistical models like Gaussian Mixture Models (GMMs) [1] and i-vectors [2] to achieve good performance. With the advance in deep learning, deep neural networks, such as TDNNs [3], ResNets [4], and ECAPA-TDNNs [5], have been prevailing for speaker embedding. Notably, the ECAPA-TDNN has achieved state-of-the-art performance on various datasets, demonstrating its superiority in speaker verification tasks.

Conventional SV systems are usually trained on "clean" utterances and perform well on near-field speech signals. Under far-field conditions, however, due to uncontrollable noise and reverberation, a severe mismatch occurs between the near-field and far-field acoustics, and these systems suffer greatly [6]. Developing an SV system that can address the adverse conditions in the far field is essential.

Researchers attempted to address the far-field challenge by modifying the system architecture, exploring adversarial learning techniques, and leveraging advanced data augmentation strategies. For instance, the author in [7] introduced the channel-interdependence enhanced Res2Net (CE-Res2Net) to aggregate speaker information from multi-scale frame-level representations and achieved performance gains on VOiCES Challenge 2019 data. The authors in [8] used a domain separation network to disentangle and suppress the domain-specific information related to far-field noise and reverberation. In [9], a population-based searching strategy was proposed to optimize the augmentation parameters and greatly boosted far-field SV performance.

On the other hand, studies have shown that text-independent SV systems can be enhanced by incorporating phonetic information into speaker representation learning. In [10], the authors adopted a multi-task learning strategy by combining a phone classifier with a speaker classifier for speaker embedding and obtained superior performance. The authors of [11] investigated the usefulness of phonetic information at the segment and frame levels. They concluded that although phonetic content at the segment (embedding) level is detrimental to SV performance, using phonetic information at the frame level is beneficial. One possible explanation for the performance improvement in [10], [11] is that shared spectral dynamics exist at the lower (frame-level) layers, which are useful for speech and speaker recognition. Enriching content information at the frame-level layers also strengthens the information essential for speaker discrimination.

The aforementioned studies only focus on the contribution of phonetic information to near-field speaker verification tasks, and they rely on phonetic labels when incorporating phonetic information. However, obtaining phonetic labels for SV is challenging because most speech-to-text systems can only output word sequences instead of phone sequences, and the phones are not time-aligned after converting from words to phones.

Inspired by the above observations, we propose a framework that can jointly train a model to perform phonetic matching and speaker verification without any phonetic labels. The framework comprises a phonetic matching component for phonetic information extraction and a speaker identification component to enforce the segment-level layers to produce speaker discriminative vectors. We refer to the framework as **J**ointly **o**ptimized **s**peaker-**e**mbedding and **ph**onetic-matching (**Joseph**). Unlike

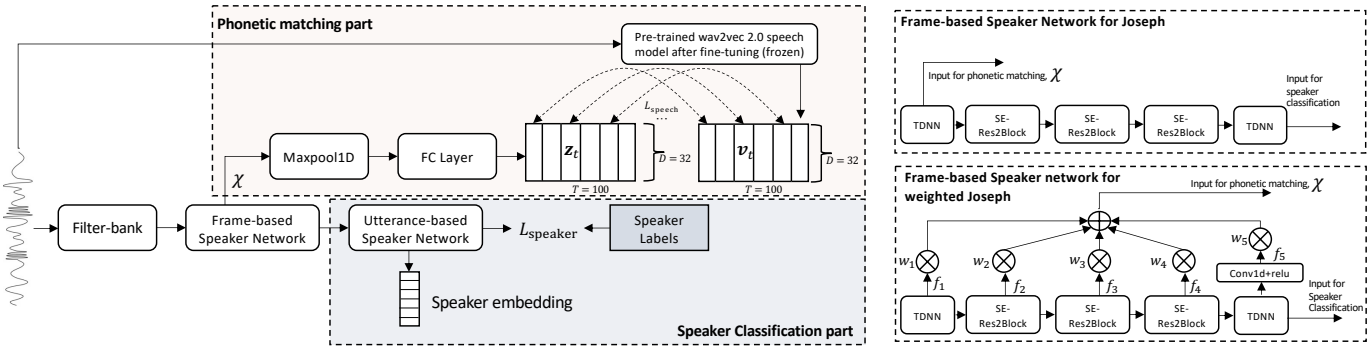

Fig. 1: *Left*: Framework of **J**ointly **o**ptimized **s**peaker-**e**mbedding and **ph**onetic-matching (Joseph) and Weighted Joseph. The utterance-based speaker network in the speaker classification part comprises a pooling layer and a fully connected layer. *Right*: The difference between Weighted Joseph and Joseph lies in the frame-based network. "⊗" represents multiplication and "⊕" represents element-wise addition.

[10], [11], phonetic labels are not required in Joseph. Instead, we use a pre-trained wav2vec 2.0 model to extract the phonetic content in an unsupervised manner. This strategy greatly saves the effort to transcribe the speech files in speaker recognition corpora. The rationale behind Joseph is that although noise and reverberation can blur speaker information in speech signals, the phonetic information extracted from wav2vec 2.0 assists in preserving the underlying acoustic dynamics shared by the speaker identity. Therefore, the degradation due to the far-field conditions can be compensated somewhat. As different frame-level layers contain distinct speaker and phonetic information, Joseph should be able to utilize the phonetic information from different frame-level layers rather than solely relying on the last layer. To address this issue, we propose Weighted Joseph, which utilizes learnable weights to aggregate the outputs of all frame-level layers as the input to the phonetic matching component of Joseph. In this way, the model can autonomously determine which frame-level layer's information is more important. Our main contributions are as follows:

1) We proposed a phonetic-aware framework called Joseph, which improves the robustness of far-field SV by exploiting phonetic information.
2) We incorporated a pre-trained wav2vec 2.0 model (after fine-tuning) in the phonetic matching part, eliminating the need for manually transcribing speaker verification datasets.
3) We improved the Joseph framework by introducing the Weighted Joseph framework, which significantly boosts the system's performance.

The rest of the paper is organized as follows. Section II introduces the wav2vec 2.0 and details the Joseph framework. Section III presents the experimental settings, and Section IV shows the results and analyses. We draw a conclusion in Section V.

II. Methodology

This section introduces the Joseph framework and its two components: phonetic matching and speaker classification.

A. Speech Recognition Model

In Fig. 1, we utilize a wav2vec 2.0 [12] network fine-tuned by CTC loss [13] as the speech model. Wav2vec 2.0 is a self-supervised learning framework that leverages a large amount of unlabeled data to learn speech representations. It takes in a waveform and produces context representations through a stack of CNN layers and Transformer layers. Through contrastive learning, the model is able to extract compact and meaningful speech representations that can be used for downstream speech tasks. Recently, pre-trained wav2vec 2.0 models have gained popularity as a front-end feature extractor in various speech applications.

B. Joseph Framework

As shown in Fig. 1, the phonetic matching part and the speaker classification part share the frame-level layers. The representations outputted from an intermediate frame-level layer are fed into the phonetic matching part. We denote these representations as $\mathcal{X} = \{\mathbf{x}_t \in \mathbb{R}^D; t = 1, \ldots, T\}$, where $\mathbf{x}_t$ is a D-dimensional vector at the t-th frame. For the speaker classification part, the feature maps produced from the last frame-level layer are processed by a pooling layer and a fully connected (FC) layer to derive an utterance-level embedding $\mathbf{e}$. The AAMSoftmax [14] loss is employed as the loss function (L_{speaker} in Fig 1).

For the phonetic matching part, the waveform is fed into the speech model and we obtain a sequence of T frames $\mathcal{V} = \{\mathbf{v}_t \in \mathbb{R}^{\widetilde{D}}; t = 1, \ldots, T\}$, where $\widetilde{D}$ is the dimension of speech vectors. A max-pooling layer is applied to $\mathcal{X}$ to ensure that the resulting $\mathcal{Z} = \{\mathbf{z}_t \in \mathbb{R}^{\widetilde{D}}; t = 1, \ldots, T\}$ has the same length as $\mathcal{V}$. We compute the speech loss as the cosine similarity between $\mathcal{Z}$ and $\mathcal{V}$:

$$L_{\text{speech}} = 1 - \frac{1}{T}\sum_{t=1}^{T}\cos(\mathbf{z}_t, \mathbf{v}_t). \tag{1}$$

Then, we average the L_{speech} across the utterances in a mini-batch. By making $\mathcal{Z}$ close to $\mathcal{V}$, we enable the frame-level lay-

ers of the speaker encoder to preserve useful phonetic information. Because phonetic information contains speaker-dependent acoustic dynamics, maintaining phonetic information at the frame level will also preserve speaker information in the embedding network. As will be demonstrated in Section IV-A, this speaker information preservation helps compensate for the performance degradation caused by far-field environments.

The total loss is defined as follows:

$$L_{\text{total}} = L_{\text{speaker}} + \lambda L_{\text{speech}}, \tag{2}$$

where L_{speaker} is the AAMSoftmax loss defined in [14] and λ is a hyperparameter that controls the contribution of phonetic information. During training, we freeze the parameters of the speech model.

C. Weighted-Joseph Framework

As shown in Fig. 1, the Joseph framework feeds a specific frame level layer's output into the phonetic matching part. However, each frame-level layer contains different phonetic and speaker information. This means that Joseph does not consider the information from other frame-level layers during the phonetic matching process. To address this issue, we have improved the Joseph framework and introduced the Weighted-Joseph (W-Joseph for short) framework.

As shown on the right side in Fig. 1, W-Joseph multiplies the output $\mathcal{F}_i = \{\mathbf{f}_{i,1}, \ldots, \mathbf{f}_{i,T}\}$ of each frame-level layer with the corresponding learnable weights w_i, where $i = \{1, \ldots, 5\}$ represents the index of the frame-level layers. To avoid dimension mismatch during element-wise addition, the frame-level output of each layer undergoes a one-dimensional convolution to unify the dimensions. After element-wise addition, we obtain $\mathcal{X}$, which is then fed into the phonetic matching part. The following equation is defined:

$$\mathcal{X} = \sum_{i=1}^{5} w_i \times \mathcal{F}_i. \tag{3}$$

In this way, Weighted Joseph considers the information from each frame-level layer in the phonetic matching part, and through a set of learnable weights w_i, the model can autonomously determine the significance of each frame-level layer's information.

III. Experimental Setup

A. Datasets and Data Preparation

The training data comprise the VoxCeleb1 development set and the VoxCeleb2 development set [15] [16], which consist of a total of 7,205 speakers. Voice activity detection (VAD) was not used. We followed the data augmentation strategy in Kaldi's recipes [17]. We added noise, music, and babble to the training data using MUSAN [18] and created reverberated speech data based on RIR [19]. For evaluation, we used the VOiCES Challenge 2019 evaluation (VOiCES19-eval) dataset [20]. The Voxceleb1 test Original (Vox-O), which comprises 40 speakers, was also used as the evaluation set.

B. Network Training

We used the standard x-vector [21] and ECAPA-TDNN [5] as our backbones. The channel size of ECAPA-TDNN is 512. The dimension of speaker embeddings is 192 for ECAPA-TDNN and 512 for x-vector, respectively. For the speech model, we used the wav2vec 2.0 model fine-tuned on the LibriSpeech dataset. The output of the wav2vec 2.0 was obtained from the projection layer of the fine-tuned model. For the unweighted Joseph, the frame-level representation from the second lowest-level TDNN of the x-vector network and the lowest-level ECAPA-TDNN were used as the input to the phonetic matching part (see the top-right of Fig. 1). For the Weighted-Joseph, the feature maps outputted by the five frame-level layers of these networks are linearly combined, as shown in Eq. 3 and the bottom-right of Fig. 1.

For ECAPA-TDNN, we extracted 80-dimensional filter-bank (Fbank) features from 16kHz audio signals using a 25ms window with a 10ms frameshift. For the x-vector network, we extracted 40-dimensional Fbank features. Each training segment in the mini-batch has a duration of 2 seconds. The batch size was set to 100 for ECAPA-TDNN and 50 for x-vector, respectively. We used an Adam optimizer with an initial learning rate of 0.001 and employed a step learning rate scheduler. The total number of epochs is 80. For the AAMSoftmax loss function, the margin is 0.2 and the scale is 30.

C. Performance Evaluation

We used a cosine scoring backend in all experiments. When performing evaluation on the Vox-O test set, we followed the setting in [5] to apply the AS-norm [22] on the scores. The performance metrics include equal error rate (EER) and minimum detection cost function (minDCF) with $P_{\text{target}} = 0.01$.

IV. Results and Analyses

We report the performance of Joseph and W-Joseph in this section. The comparison with conventional speaker embeddings is detailed.

A. Main Results

Table I presents the results of the baselines, Joseph, and W-Joseph on the VOiCES19-eval, Vox-O (clean), and Vox-O (noise) datasets. The baselines use the x-vector network or ECAPA-TDNN without phonetic matching. From Table I, it is evident that Joseph and W-Joseph outperform the baselines for both x-vector and ECAPA-TDNN on all datasets. In particular, on VOiCES19-eval, for ECAPA-TDNN, W-Joseph reduces the EER by 14.2% and minDCF by 18.69%. For x-vector, W-Joseph achieves a reduction of 22.99% and 33.44% on EER and minDCF, respectively. These reductions demonstrate the effectiveness of Joseph and W-Joseph in leveraging phonetic information for far-field SV.

To verify that Joseph and W-Joseph can partially compensate for the performance degradation due to adverse conditions under the far-field scenarios, we investigated their performance

TABLE I: Comparison of our methods (Joseph and W-Joseph) with the baselines (without phonetic matching) on the clean and noisy Vox-O datasets and the VOiCES19-eval sets.

Row	System	Speaker Embedding Network	VOiCES19-eval EER (%)	VOiCES19-eval minDCF	Vox-O (clean) EER (%)	Vox-O (clean) minDCF	Vox-O (noise) EER (%)	Vox-O (noise) minDCF
1	Baseline 1	x-vector	7.81	0.598	2.23	0.219	5.82	0.470
2	Joseph (Proposed)	x-vector	6.43	0.433	2.16	0.205	5.14	0.446
3	W-Joseph (Proposed)	x-vector	**6.01**	**0.398**	**2.01**	**0.174**	**4.88**	**0.173**
4	Baseline 2	ECAPA-TDNN	5.79	0.428	1.19	0.165	3.95	0.380
5	Joseph (Proposed)	ECAPA-TDNN	5.13	0.374	1.10	0.136	3.31	0.311
6	W-Joseph (Proposed)	ECAPA-TDNN	**4.97**	**0.349**	**1.05**	**0.143**	**3.05**	**0.252**

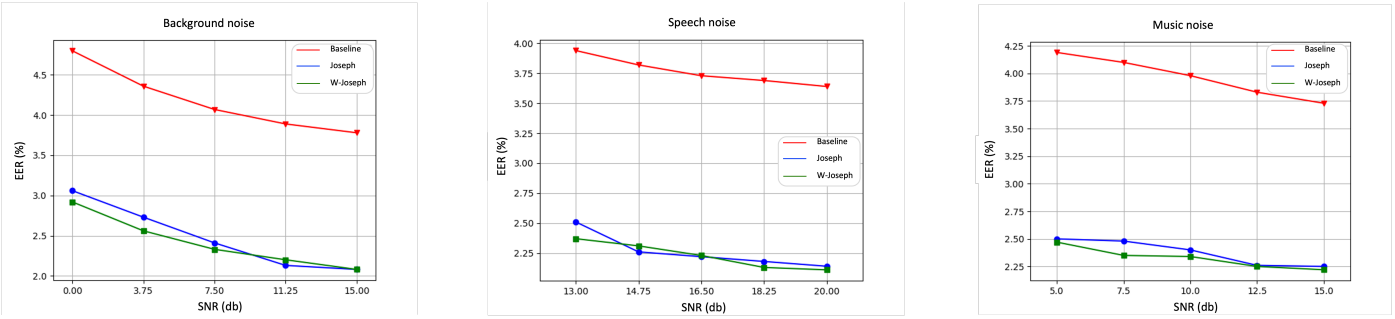

Fig. 2: The impact of different noise types in MUSAN and signal-to-noise (SNR) on the baseline, Joseph, and W-Joseph. The results were based on Vox-O (noise). The speaker embedding network is ECAPA-TDNN.

TABLE II: Comparison of our method (Joseph and W-Joseph) with other methods on the VOiCES19-eval dataset. The best results for each model are highlighted in bold.

System	Speaker Embedding	EER (%)	minDCF
[7]	CE-ResNet	5.72	0.423
[23]	x-vector	8.55	0.552
[24]	x-vector	6.85	0.536
[8]	x-vector	6.51	0.6309
[25]	x-vector	6.42	0.501
[26]	ECAPA-TDNN	5.90	_
Joseph (Proposed)	x-vector	6.43	0.433
W-Joseph (Proposed)	x-vector	**6.01**	**0.398**
Joseph (Proposed)	ECAPA-TDNN	5.13	0.374
W-Joseph (Proposed)	ECAPA-TDNN	**4.97**	**0.349**

TABLE III: Impact of preserving phonetic information at different frame-level layers of an ECAPA-TDNN on Joseph.

Frame-level Layer	Network Block	EER (%)	minDCF
Layer 4	TDNN	5.28	0.392
Layer 3	SE-Res2Block	5.54	0.379
Layer 2	SE-Res2Block	5.33	0.385
Layer 1	SE-Res2Block	5.27	0.390
Layer 0	TDNN	**5.13**	**0.374**

TABLE IV: Impact of preserving phonetic information at different frame-level layers of an x-vector network on Joseph.

Frame-level Layer	Network Block	EER (%)	minDCF
Layer 4	TDNN	6.85	0.483
Layer 3	TDNN	6.74	0.477
Layer 2	TDNN	6.54	0.485
Layer 1	TDNN	**6.43**	**0.433**
Layer 0	TDNN	6.44	0.459

on the clean and noisy Vox-O datasets. The "clean" set refers to the standard Vox-O test data, and the noisy Vox-O set was created by randomly adding noise and reverberation to the standard (clean) Vox-O data, following the data augmentation strategy in Section III-B. From Table I, we observe that Joseph and W-Joseph outperform the baselines on both clean and noisy Vox-O datasets. On the clean Vox-O, our methods achieve a slight improvement over the baselines. This improvement confirms the conclusion in [10], [11] that using phonetic content can benefit text-independent SV. On the noisy Vox-O dataset, all systems exhibit substantial performance degradation compared with the clean counterpart. Nevertheless, phonetic-aware systems (Joseph and W-Joseph) achieve remarkably greater performance gain over the baseline systems. This observation verifies our motivation that incorporating phonetic information into the speaker embedding system can improve SV performance, particularly in far-field

TABLE V: Impact of λ (in eq. 2) on the Joseph framework. The best results are highlighted in bold.

Speaker embedding	λ	EER (%)	minDCF
ECAPA-TDNN	0.001	5.69	0.410
	0.004	5.67	0.418
	0.01	5.59	0.401
	0.1	**5.13**	**0.374**
	0.4	5.85	0.448

TABLE VI: Impact of different speech models on the W-Joseph framework. The results based on the VOiCEs19-eval set.

Speaker embedding	Speech Model	EER (%)	minDCF
ECAPA-TDNN	Wav2vec2	4.97	0.349
	Hubert	5.04	0.343
	WavLM	**4.68**	**0.321**

environments with noise and reverberation.

To further demonstrate the effectiveness of our methods, we compared our method with several recent approaches on the VOiCES19-eval dataset. Table II shows the results. We observed that both Joseph and W-Joseph outperform other methods. W-Joseph achieves better performance compared to Joseph because it focuses on the phonetic information at every frame-level layer and allows the model to autonomously determine which layer's phonetic information is more important through learnable weights.

We have also conducted additional experiments on the Vox-O test set to evaluate the effects of three distinct noise categories in MUSAN: speech, noise, and music, at different SNR levels. Fig. 2 shows the results. From Fig. 2, it is evident that the phonetic aware systems demonstrate superior performance compared to the baseline across different types of noise and SNR levels. Furthermore, When the SNR decreases, the Joseph and W-Joseph system exhibits a relatively smaller decline in performance. This finding indicates the robustness of our methods to noise and reverberation, highlighting its ability to maintain stable performance even in challenging acoustic environments.

B. Ablation Study

We argue that the lower frame-level features contain richer shared spectral dynamic information related to speech and speakers. To support this argument, we conducted ablation studies to show the impact of preserving phonetic information at different frame-level layers of Joseph. Table III and Table IV show the results. In Table III, Layer 0 and Layer 4 correspond to the initial and final TDNN layers of the ECAPA-TDNN, respectively. The remaining three layers correspond to the three SE-Res2Blocks. In Table IV, Layer 0 to Layer 4 correspond to the five TDNN layers of the x-vector network.

Table III shows that the performance improvement of Joseph becomes more prominent when we feed features from lower layers into the phonetic matching part. Specifically, when we present features from the initial TDNN layer (Layer 0) to the phonetic matching part, we obtained the best result with an EER of 5.13%. However, performance gradually drops when we preserve phonetic information at the upper layers (with more abstract representations). This result is reasonable because the lower-level feature maps contain more speaker and content information that is entangled together. By contrast, the representations at the upper layers are more speaker-specific. Therefore, it is preferable to exploit phonetic information at lower layers. In Table IV, feeding the output of the bottom frame-level layers into the phonetic matching part yields better results. However, it is not the lowest layer. Our analysis suggests that the TDNN layers do not possess the same level of ability as the SE-Res2Blocks to filter out speaker-independent information. This is also the reason for using the bottom layer for phonetic information extraction in Section IV-B.

We also investigated the effect of λ in Eq. 2 on Joseph. The results are shown in Table V. We observe that the best performance is achieved when $\lambda = 0.1$, with an EER of 5.13%. As λ increases, the performance of Joseph gradually deteriorates. When λ was set to 0.4, the EER of the Joseph system is higher than that of Baseline 2 in Table I. The above observations suggest that excessive phonetic information can cause the speaker embedding network to focus on the content details and pay less attention to the speaker information, leading to performance degradation.

Table VI demonstrates the impact of different speech models on W-Joseph. It shows that W-Joseph performs the best when the speech model is WavLM.

V. Conclusions

In this paper, we propose two phonetic aware systems (Joseph and W-Joseph) to improve far-field SV performance. By using a pre-trained speech recognition model, we incorporate the phonetic information into the conventional speaker encoders. Also, we eliminate the reliance on transcriptions for the speech recognition task. Experimental results demonstrated that leveraging phonetic information can improve the performance of far-field speaker verification. In the future, we plan to replace the speech model with models suitable for other languages to test their performance on a wider variety of far-field datasets in different languages.

References

[1] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1-3):19–41, 2000.

[2] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2010.

[3] Alexander Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J Lang. Phoneme recognition using time-delay neural networks. In *Backpropagation*, pages 35–61. Psychology Press, 2013.

[4] Hossein Zeinali, Shuai Wang, Anna Silnova, Pavel Matějka, and Oldřich Plchot. BUT system description to voxceleb speaker recognition challenge 2019. *arXiv preprint arXiv:1910.12592*, 2019.

[5] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. *in Proc. Interspeech*, pages 3830–3834, 2020.

[6] Qin Jin, Tanja Schultz, and Alex Waibel. Far-field speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2023–2032, 2007.

[7] Ling-jun Zhao and Man-Wai Mak. Channel interdependence enhanced speaker embeddings for far-field speaker verification. In *12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2021.

[8] Lu Yi and Man-Wai Mak. Adversarial separation and adaptation network for far-field speaker verification. In *Proc. Interspeech*, pages 4298–4302, 2020.

[9] Weiwei Lin and Man-Wai Mak. Robust speaker verification using population-based data augmentation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7642–7646. IEEE, 2022.

[10] Yi Liu, Liang He, Jia Liu, and Michael T Johnson. Speaker embedding extraction with phonetic information. *arXiv preprint arXiv:1804.04862*, 2018.

[11] Shuai Wang, Johan Rohdin, Lukás Burget, Oldrich Plchot, Yanmin Qian, Kai Yu, and Jan Cernockỳ. On the usage of phonetic information for text-independent speaker embedding extraction. In *Proc. Interspeech*, pages 1148–1152, 2019.

[12] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

[13] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International conference on Machine Learning*, pages 369–376, 2006.

[14] Xu Xiang, Shuai Wang, Houjun Huang, Yanmin Qian, and Kai Yu. Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition. In *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1652–1656, 2019.

[15] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.

[16] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.

[17] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The Kaldi speech recognition toolkit. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[18] David Snyder, Guoguo Chen, and Daniel Povey. MUSAN: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*, 2015.

[19] Marco Jeub, Magnus Schafer, and Peter Vary. A binaural room impulse response database for the evaluation of dereverberation algorithms. In *Proc. 16th International Conference on Digital Signal Processing*, pages 1–5, 2009.

[20] Mahesh Kumar Nandwana, Julien Van Hout, Mitchell McLaren, Colleen Richey, Aaron Lawson, and Maria Alejandra Barrios. The voices from a distance challenge 2019 evaluation plan. *arXiv preprint arXiv:1902.10828*, 2019.

[21] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust DNN embeddings for speaker recognition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333, 2018.

[22] Pavel Matejka, Ondrej Novotnỳ, Oldrich Plchot, Lukas Burget, Mireia Diez Sánchez, and Jan Cernockỳ. Analysis of score normalization in multilingual speaker recognition. In *Proc. Interspeech*, pages 1567–1571, 2017.

[23] Sergey Novoselov, Aleksei Gusev, Artem Ivanov, Timur Pekhovsky, Andrey Shulipa, Galina Lavrentyeva, Vladimir Volokhov, and Alexandr Kozlov. STC speaker recognition systems for the VOiCES from a distance challenge. *arXiv preprint arXiv:1904.06093*, 2019.

[24] Zhor Benhafid, Sid Ahmed Selouani, Mohammed Sidi Yakoub, and Abderrahmane Amrouche. Residual time-restricted self-attentive TDNN speaker embedding for noisy and far-field conditions. In *Proc. 2022 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 471–476, 2022.

[25] Weiwei Lin and Man-Wai Mak. Mixture representation learning for deep speaker embedding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:968–978, 2022.

[26] Sandipana Dowerah, Romain Serizel, Denis Jouvet, Mohammad Mohammadamini, and Driss Matrouf. Joint optimization of diffusion probabilistic-based multichannel speech enhancement with far-field speaker verification. In *Proc. IEEE Spoken Language Technology Workshop (SLT)*, pages 428–435, 2023.